主流 AI 模型 API 平台全景指南：注册、模型、价格与选型建议（2025 年版）

作者：阮小燕
面向人群：在校大学生、AI 开发者、开源贡献者、AI 科技博主
更新日期：2025 年 11 月 12 日

随着大语言模型（LLM）生态日益繁荣，开发者已不再局限于 OpenAI 的 GPT 系列。国际上，Anthropic、Google、Meta、Mistral 等厂商纷纷开放 API，甚至出现了一批聚合多模型的"路由平台"；国内，阿里云、腾讯、百度、智谱、月之暗面等科技巨头和创业公司也推出了强大的中文大模型服务。面对数十种选择，如何快速了解各平台特点、完成注册、评估成本与适用场景，成为 AI 开发的第一步。

本文将系统梳理 **21 个主流 AI API 平台**（13 个国际平台 + 8 个中国平台），涵盖：

- 平台定位与核心优势
- 支持的代表性模型
- 注册与获取 API Key 的完整流程
- 免费额度与计费策略
- 适合的开发场景建议

助你在项目初期做出明智选型。

目录

16. 百度千帆（文心一言）
17. 智谱 AI（ChatGLM）
18. 月之暗面（Kimi）
19. MiniMax
20. 字节豆包
21. 讯飞星火

第三部分：对比与选型

- 国际平台横向对比表
- 中国平台横向对比表
- 选型建议

第一部分：国际 AI 平台

1. OpenAI —— 商业化 AI 的奠基者

核心模型

- **GPT-5**（2025年8月发布，最新旗舰模型，统一系统）
  - GPT-5-Low/High：可快速响应或深度思考
  - AIME 2025数学94.6%，SWE-bench编程74.9%，MMMU多模态84.2%
  - 降低45%事实错误率（相比GPT-4o），思考模式降低80%错误率
- **GPT-4.5**（2025年2月发布，过渡模型）
- **GPT-4o**（多模态、低延迟）
- **GPT-4o mini**（轻量版，有免费层）
- **o1/o3 系列**（深度推理优化）

注册流程

1. 访问 https://platform.openai.com
2. 使用邮箱或 Google 账号注册
3. **绑定信用卡**（即使使用免费试用也需验证）
4. 进入 API Keys 页面 创建密钥

> ⚠️ 注意：中国内地手机号无法完成短信验证，建议使用海外手机号或教育邮箱。

免费额度与价格

- **免费额度**：GPT-4o mini 提供免费层（无需信用卡）
- **GPT-5**：输入 $5/1Mtokens$， 输出15/1M tokens（Pro用户更多配额）
- **GPT-4o**：输入 $5/1Mtokens$， 输出15/1M tokens
- **GPT-4o mini**：输入 $0.15/1Mtokens$， 输出0.6/1M tokens
- **o1/o3 系列**：价格较高（约 $15–60$/1M tokens）

适合场景

- 商业产品开发

- 对稳定性和文档质量要求高的项目
- 需要多模态（图像+文本）能力的应用

2. Anthropic —— 安全优先的高质量模型提供商

核心模型

- **Claude Sonnet 4.5**（2025年9月29日发布，最新！）
  - 世界最佳编程模型
  - 构建复杂Agent的最强模型
  - 电脑使用能力最佳，推理和数学能力显著提升
- **Claude Opus 4.1**（2025年8月5日发布）
  - Agent任务、真实编程和推理优化
- **Claude Haiku 4.5**（2025年10月15日发布）
  - 小而快速，低延迟优化
- **Claude Sonnet 4** & **Claude Opus 4**（2025年5月22日，Claude 4家族）

注册流程

1. 访问 https://console.anthropic.com
2. 注册账号并完成邮箱验证
3. 部分区域需**申请 API 访问权限**（1–3 个工作日）
4. 审核通过后，在 API Keys 页面 创建 Key

免费额度与价格

- **免费额度**：新用户 $5 免费额度（无需信用卡）
- **Claude Sonnet 4.5**：输入 $3/1Mtokens$，输出15/1M tokens
- **Claude Opus 4.1**：输入 $15/1Mtokens$，输出75/1M tokens
- **Claude Haiku 4.5**：输入 $0.25/1Mtokens$，输出1.25/1M tokens（性价比极高）
- **Claude Sonnet 4 / Opus 4**：与4.1系列类似定价

适合场景

- 高质量长文本生成（如论文、报告）
- 金融、法律等对输出安全性要求高的领域
- 需要 200K+ 上下文的复杂推理任务

3. Google AI (Gemini) —— 深度集成 Google 生态

核心模型

- **Gemini 2.5 Pro**（2025年当前最强，LMArena排名#1）
  - 思考模型，推理前进行深度思考
  - 数学与科学基准领先（GPQA、AIME 2025）
  - SWE-Bench Verified得分63.8%
  - Humanity's Last Exam无工具18.8%

- **Gemini 2.5 Flash / Flash-Lite**（2025年9月更新，更高质量和效率）
- **Gemini 3.0**（即将发布，2025年11月预计）
  - 1M上下文限制
  - 更强大的AI Agent能力
  - 增强多模态、编程能力

注册流程

1. 访问 https://aistudio.google.com 或 Vertex AI Console
2. 使用 Google 账号登录
3. 启用 **Gemini API**（需在 Cloud Console 中启用 billing）
4. 创建 **API Key** 或使用 **Service Account**（推荐后者用于生产）

> 💡 小技巧：Google AI Studio 提供"临时聊天"功能，无需配置即可测试模型。

免费额度与价格

- **免费层**：5-15 RPM（每分钟请求数），允许商业使用，100万token上下文可用
- **新用户**：$300 Google Cloud 信用（有效期 90 天）
- **Gemini 2.5 Pro**：输入 $3.5/1Mtokens$，输出10.5/1M tokens
- **Gemini 2.5 Flash**：输入 $0.35/1M tokens（性价比极佳）
- **Gemini 3.0**：价格待公布（即将发布）

适合场景

- 与 Google Workspace（Docs、Gmail）深度集成
- 处理超长文档（1M tokens 支持）
- 图像+文本多模态任务（如 PDF 解析）

4. Together AI —— 开源模型的首选推理平台

核心模型

- **Llama 4 系列**（2025年4月发布，最新！）
  - **Llama 4 Scout**：17B激活参数（16专家），10M tokens上下文
  - **Llama 4 Maverick**：17B激活参数（128专家），击败GPT-4o和Gemini 2.0 Flash
  - **Llama 4 Behemoth**：训练中，超越GPT-4.5和Claude Sonnet 3.7
- **Llama 3.1 405B / 70B / 8B**（上一代）
- **Mixtral 8x22B**
- **Qwen 2.5 / 3, Yi 1.5, DeepSeek-Coder**
- 支持自定义微调模型部署

注册流程

1. 访问 https://together.ai
2. 使用 GitHub 或邮箱注册
3. 登录后自动获得 **$25 免费额度**
4. 进入 API Keys 页面 创建 Key

免费额度与价格

- **免费额度**：$25（自动发放）
- **Llama 3 8B Lite**：$0.10/1M tokens（比 GPT-4o-mini 便宜 6 倍）
- **Llama 3.1 70B**：$0.8/1M tokens
- **Llama 3.1 405B**：约 $3/1M tokens
- 支持按秒计费（适合间歇性调用）

适合场景

- 开源模型性能对比实验
- 学术研究、课程项目
- 需要长上下文（128K+）且控制成本

5. Fireworks AI —— 高性能推理与企业级 SLA

核心模型

- **Llama 4 Scout / Maverick**（2025年4月首发，最快Llama 4 API！）
  - Llama 4 Maverick：400B总参数，17B激活参数，128专家
  - 吞吐量：145 tokens/秒（H200）
  - 100万token上下文窗口
  - 原生多模态（文本+图像理解）
- **Llama 3.1 系列（官方优化版）**
- **Mixtral, Phi-3, Code Llama**
- 自研 **Firefunction V2**（函数调用优化）

注册流程

1. 访问 https://fireworks.ai
2. 注册账号（支持 GitHub / Google）
3. 自动发放 **$1 免费额度**
4. 进入 API Keys 页面 创建 Key

免费额度与价格

- **免费额度**：$1（无需绑卡）
- **Llama 3.1 70B**：$0.9/1M tokens
- **支持 P99 延迟 <500ms** 的 SLA 保障
- 提供专用实例（Dedicated Endpoints）用于高并发

适合场景

- 需要最快Llama 4推理速度的应用
- 低延迟要求的实时应用（如聊天机器人）
- 生产环境部署（P99延迟<500ms SLA保障）
- 需要函数调用（function calling）高精度的项目
- 多模态应用（文本+图像处理）

6. Perplexity —— 联网推理与研究导向

核心模型

- **Sonar (Small / Medium / Large)**：联网实时回答
- **支持调用最新模型**（通过统一接口）：
  - **GPT-5** (OpenAI 2025年8月)
  - **Claude Sonnet 4.5** (Anthropic 2025年9月)
  - **Llama 4 Scout / Maverick** (Meta 2025年4月)
  - 以及 Gemini 2.5 等主流模型

注册流程

1. 访问 https://www.perplexity.ai
2. 使用 Google 或邮箱注册
3. API 访问在 Perplexity API 页面 获取
4. Pro 用户包含 $5/月 API 信用

免费额度与价格

- **免费额度**：Pro 用户包含 $5/月 API 信用
- **Sonar Small**：$0.2/1M tokens
- **Sonar Large**：$5/1M tokens
- 联网查询不额外收费

适合场景

- 实时信息检索（如新闻、股价、学术进展）
- 构建"会联网"的 AI 助手
- 学术研究型问答系统

7. Mistral AI —— 欧洲轻量高效代表

核心模型

- **Magistral Small / Medium**（2025年6月，推理模型）
  - Chain-of-thought推理能力
  - Magistral Small开源
- **Mistral Small 3.2**（2025年6月）
  - 改进指令遵循、输出稳定性、函数调用
- **Mistral Medium 3**（2025年5月7日）
- **Mistral Large 2.1 / 24.11**
- **Codestral 25.01**（代码生成专用）
- **Mistral Small 3.1**（2025年3月，轻量高效）
- **Pixtral Large**（多模态）

注册流程

1. 访问 https://console.mistral.ai

2. 注册账号（需邮箱验证）
3. 在 API Keys 页面 创建 Key
4. 可选绑定支付方式（支持 PayPal）

免费额度与价格

- **免费额度**：$500 免费 API 信用（新用户）
- **Pixtral 12B**：免费（在 le Chat 和 API 中）
- **Mistral Large 2**：输入 $2/1Mtokens$，输出6/1M tokens
- **Codestral**：价格下降 80%，现为 $0.2/1M tokens
- **Mistral Nemo**：价格下降 50%

适合场景

- 欧洲开发者（数据合规）
- 代码生成与补全
- 资源受限设备的轻量推理
- 预算有限的原型开发（免费额度慷慨）

8. OpenRouter —— 多模型统一接入层（聚合平台）

核心特点

- 聚合所有主流平台的**最新模型**（300+ 模型）：
  - **GPT-5**（OpenAI 2025年8月）
  - **Claude 4 系列**（Sonnet 4.5, Opus 4.1, Haiku 4.5）
  - **Llama 4**（Scout, Maverick）
  - **Gemini 2.5 Pro**（Google）
  - **Qwen3-Max, ERNIE 5.0** 等中国模型
- 统一 API 接口，一次集成即可切换模型
- 支持自定义 HTTP-Referer 和 X-Title，鼓励开源透明
- 支持 BYOK（Bring Your Own Key）：首 100 万请求/月免费

注册流程

1. 访问 https://openrouter.ai
2. 使用 GitHub / Google / Apple 登录
3. 进入 Keys 页面 创建 API Key
4. （可选）绑定信用卡启用付费

免费额度与价格

- **免费额度**：$1 试用额度
- **BYOK**：首 100 万请求/月免费，之后收取底层成本的 5%
- **直接使用**：价格 = 底层模型价格 + 5.5% 平台服务费（最低 $0.80）
- 加密货币支付收取 5% 手续费
- 无隐藏费用，用量透明

适合场景

- 快速原型开发与模型对比
- 开源项目（支持 Referer 标注）
- 不想管理多个 API Key 的轻量项目
- 探索 300+ 模型的开发者

9. Groq —— 超高速推理引擎

核心特点

- 基于自研 **LPU（Language Processing Unit）**芯片
- **世界最快推理速度**：
  - Llama 4 达到 **625 tokens/秒**（业界最快！）
  - SRAM带宽 80TB/秒，比GPU HBM快10倍
- Meta官方Llama 4 API推理引擎（2025年4月合作）
- IBM战略合作伙伴（2025年10月）：速度快5倍，成本更低

核心模型

- **Llama 4 Scout / Maverick**（2025年4月发布当日上线！）
  - 400B总参数，17B激活参数
  - 吞吐量：625 tokens/秒
- **Llama 3.1 405B / 70B / 8B**
- **Mixtral 8x7B**
- **Gemma 7B**

注册流程

1. 访问 https://console.groq.com
2. 注册账号
3. 在控制台获取免费 API Key

免费额度与价格

- **免费层**：提供免费 API 访问（有速率限制）
- **Developer Tier**：按 token 付费，价格竞争力强
- **Enterprise Tier**：定制化解决方案

适合场景

- **需要极速推理的应用**（625 tokens/秒，业界最快）
- 需要极低延迟的实时应用（聊天机器人、实时翻译）
- 大规模推理任务（企业级部署）
- 成本敏感的生产环境（比GPU便宜且快5倍）
- Llama 4多模态应用

10. Replicate —— 按需运行 AI 模型

核心特点

- 运行 AI 模型的云平台，无需管理基础设施
- 支持文本、图像、视频、音频生成等多种AI模型
- 按实际运行时间计费

热门模型（2025年11月最新）

- **Llama 4 Scout / Maverick**（Meta 2025年4月，多模态）
- **Veo 3.1**（Google 视频生成，最新！）
  - 支持参考图像、首尾帧控制
  - 增强的图像转视频功能
- **Seedance Pro**（文本/图像转视频，480p/1080p，5s/10s）
- **SDXL**（Stable Diffusion XL图像生成）
- **Meta MusicGen**（音乐生成）
- 数字人视频生成（单图+音频+文本）

注册流程

1. 访问 https://replicate.com
2. 使用 GitHub 或邮箱注册
3. 在控制台创建 API Token

免费额度与价格

- **按时间计费**：
  - CPU：$0.0001/秒（公开模型）
  - Nvidia T4 GPU：$0.000225/秒（公开模型）
  - 8x H100 GPU：约 $0.0122/秒
- **按输出计费**（部分模型）：
  - Veo 3.1：约 $0.50/秒视频
  - Seedance Pro：约 $0.098/秒视频
  - Llama 4：按token计费（具体见模型页面）

适合场景

- 视频生成（Veo 3.1, Seedance Pro）
- 图像生成（SDXL, 数字人）
- 音频/音乐生成（MusicGen）
- 多模态应用（Llama 4 Scout/Maverick）
- 快速测试最新开源模型
- 无需部署基础设施的创意项目

11. Hugging Face —— 开源模型中心

核心特点

- **全球最大的开源 AI 模型社区**（100万+ 模型）
- 提供 Inference API 和 Inference Endpoints
- **托管所有最新开源模型**（2025年11月）：
  - **Llama 4**（Scout, Maverick, Behemoth）
  - **Qwen3**（最强开源，Apache 2.0）
  - **GLM-4.6, Hunyuan Large, Kimi K2**（中国开源）
  - **Mistral Magistral, Jamba 1.7** 等

注册流程

1. 访问 https://huggingface.co
2. 使用邮箱或 GitHub 注册
3. 在 Access Tokens 页面 创建 Token

免费额度与价格

- **免费层**：每月免费推理额度（所有用户）
- **PRO 计划**：$9/月，包含 20× 推理额度
- **Inference Endpoints**：按小时计费
  - CPU：$0.032/核心/小时
  - GPU：$0.5/小时起

适合场景

- **开源模型首选平台**（Llama 4, Qwen3, Hunyuan Large等）
- 学术研究和教学
- 需要模型微调和私有化部署的企业
- 探索和比较最新开源模型（100万+选择）
- 模型托管和分享（开发者社区）

12. Cohere —— 企业级 NLP 解决方案

核心模型（2025年最新）

- **Command A 03-2025**（2025年3月，最强！）
  - 111B参数，256K上下文
  - **150%吞吐量提升**（仅需2个GPU：A100/H100）
  - 全领域最强性能
- **Command A Translate 08-2025**（2025年8月）
  - 支持23种语言的SOTA翻译
- **Command A Reasoning**（混合推理模型）
  - 擅长复杂Agent任务
  - 支持英语+22种其他语言
- **Command R / R+ 08-2024**（上一代）

- **Embed**（文本嵌入）
- **Rerank**（搜索排序）

注册流程

1. 访问 https://cohere.com
2. 注册账号
3. 获取 Trial API Key（自动提供）

免费额度与价格

- **Trial API Key**：
  - 5,000 生成单位/月（Generate/Summarize）
  - 100 次/分钟（Embed/Classify）
  - 不可用于生产环境
- **Production API**：
  - Command R：输入 $0.15/1Mtokens$，输出0.60/1M tokens
  - Embed：$0.10/1M tokens

适合场景

- **企业级Agent应用**（Command A Reasoning）
- **多语言翻译**（23语言，SOTA性能）
- 企业级NLP应用（Command A 03-2025）
- 需要高质量文本嵌入的RAG系统（Embed）
- 搜索和推荐系统（Rerank）
- 高效推理（仅需2个GPU，150%吞吐量）

13. AI21 Labs —— 任务型 API 专家

核心模型

- **Jamba 1.7**（2025年7月，最新！）
  - 混合SSM-Transformer架构 + MoE
  - 最强长上下文模型（256K tokens）
  - 开源授权，支持企业私有化部署
- **Jamba 1.6**（2025年3月，企业版）
- **Jamba 1.5 Large / Mini**（2024年8月）
- **Jurassic-2 Ultra / Mid**（上一代）

AI编排平台

- **Maestro**（2025年11月11日发布！）
  - 智能编排层，解决Agent可靠性问题
  - 针对关键企业工作流的AI规划系统
  - 平衡准确性与自动化

注册流程

1. 访问 https://www.ai21.com
2. 注册账号
3. 获取 API Key

免费额度与价格

- **免费额度**：$10 信用（有效期 3 个月）
- **定价示例**：
  - Jurassic-2 Mid：$0.0125/1M tokens
  - Jurassic-2 Ultra：$0.0188/1M tokens

适合场景

- **超长上下文处理**（Jamba 1.7, 256K tokens）
- **企业级Agent编排**（Maestro平台）
- 需要私有化部署的企业（Jamba开源）
- 关键任务工作流（Maestro智能编排）
- 任务型API（摘要、改写、语法检查等）
- 多语言支持需求

第二部分：中国本土 AI 平台

14. 阿里云百炼（通义千问）

核心模型

- **Qwen3-Max "深度思考"模式**（2025年11月2日最新！）
  - 超过1万亿参数（1T），MoE架构
  - 预训练数据36T tokens
  - 双模式：Instruct（快速对话）+ Thinking（深度思考）
  - Thinking模式：AIME 25和HMMT达到100%准确率
- **Qwen3 系列**（2025年4月29日，世界最强开源）
  - 2个MoE模型 + 6个Dense模型
  - Apache 2.0开源
- **Qwen-Max / Plus / Turbo**（上一代）
- **Qwen-VL**（多模态视觉）

注册流程

1. 访问 https://bailian.console.aliyun.com
2. 使用阿里云账号登录（需实名认证）
3. 开通百炼服务
4. 在控制台创建 API Key

免费额度与价格

- **免费额度**：新用户赠送 100 万 tokens（主流模型各限免 100 万）
- **定价**（阶梯计费）：
  - Qwen-Max：约 ¥40/百万 tokens
  - Qwen-Plus：约 ¥4/百万 tokens
  - Qwen-Turbo：约 ¥2/百万 tokens
- **DeepSeek-R1**：新用户额外赠送 100 万 tokens

适合场景

- 中文语料处理
- 阿里云生态集成
- 电商、客服等业务场景
- 需要国内合规的企业应用

中国平台特殊优势

- 国内访问速度快，无需翻墙
- 支付宝、微信支付等本地支付方式
- 强大的中文理解和生成能力
- 符合中国数据合规要求

15. 腾讯混元

核心模型

- **Hunyuan Large**（2025年11月5日开源，最新！）
  - MoE架构
  - 完全开源，支持企业微调和部署
- **Hunyuan 3D World Model 1.0**（2025年7月27日，WAIC 2025）
  - 行业首个开源沉浸式、可交互、可模拟世界生成模型
- **Hunyuan-A13B**（2025年6月27日）
  - 首个混合推理MoE模型
  - 80B总参数，仅13B激活
  - 更快推理速度
- **Hunyuan-Pro / Standard / Lite**（基础系列）
- **Hunyuan-Standard-256K**（长文本）

注册流程

1. 访问 https://cloud.tencent.com/product/hunyuan
2. 使用腾讯云账号登录（需实名认证）
3. 开通混元服务
4. 在控制台创建密钥

免费额度与价格

- **免费额度**：新用户赠送 100 万 tokens（有效期 12 个月）

- **混元-Lite**：完全免费（上下文将升级至 256K）
- **混元-Standard**：
  - 输入：¥0.0045/千 tokens（下降 55%）
  - 输出：¥0.005/千 tokens（下降 50%）
- **混元-Pro**：
  - 输入：¥0.03/千 tokens（下降 70%）

适合场景

- 腾讯生态集成（微信、QQ、企业微信）
- 游戏、社交、内容创作
- 需要免费大模型的个人开发者（Lite 版本）
- 中小企业 AI 应用

中国平台特殊优势

- 与腾讯云服务深度集成
- 微信小程序、公众号开发
- 国内访问速度优异
- 免费模型性能出色

16. 百度千帆（文心一言）

核心模型

- **ERNIE 5.0-Preview-1022**（2025年11月8日公布，最新！）
  - LMArena文本排行榜：超越GPT-5-High、Qwen3-Max、DeepSeek-R1
  - 排名：全球第二、国内第一
  - 擅长：创意写作（排名第一）、复杂长问题理解、指令遵循
  - 即将正式发布
- **ERNIE 4.0**（文心一言旗舰版，2025年3月16日起完全免费）
- **ERNIE 3.5**（免费）
- **ERNIE Speed / Tiny**（快速/轻量级）

注册流程

1. 访问 https://cloud.baidu.com/product/wenxinworkshop
2. 使用百度账号登录（需实名认证）
3. 开通千帆服务
4. 创建应用并获取 API Key

免费额度与价格

- **重大更新**：文心一言 4.0 于 2025 年 3 月 16 日起**完全免费**（网页版和 App）
- **API 定价**：已大幅下调（比 DeepSeek R1 再降 50%）
- ERNIE 3.5：免费
- 具体 API 价格请查看官方最新文档

适合场景

- 百度生态集成（百度搜索、小程序）
- 中文内容创作、问答
- 教育、医疗等垂直领域
- 免费使用高性能中文模型

中国平台特殊优势

- 旗舰模型完全免费使用
- 强大的中文理解能力
- 与百度搜索、地图等服务集成
- 适合中小企业和个人开发者

17. 智谱 AI（ChatGLM）

核心模型

- **GLM-4.6**（2025年9月30日发布，最新！）
  - 总参数355B，激活参数32B
  - 上下文从128K提升至200K
  - 代码能力对齐Claude Sonnet 4，国产模型最强
  - 74个真实场景编程任务测试中领先
  - 与Claude Sonnet 4比肩，稳居国产模型首位
  - 已在寒武纪国产芯片上实现FP8+Int4混合量化部署
- **GLM-Experimental**（实验模型，擅长PPT制作）
- **GLM-4.5 / 4-Plus**
- **GLM-4-Flash / FlashX**（免费/超低价）
- **GLM-Z1-Air**（深度推理）
- **GLM-4-Long**（超长上下文）

注册流程

1. 访问 https://open.bigmodel.cn
2. 注册智谱账号（手机号验证）
3. 实名认证（可选，部分功能需要）
4. 在控制台创建 API Key

免费额度与价格

- **GLM-4-Flash**：完全免费
- **GLM-4-FlashX**：¥10/亿 tokens（性价比之王）
- **GLM-4.5**：输入 ¥0.8/百万 tokens，输出 ¥2/百万 tokens
- **GLM-4-Plus**：¥5/百万 tokens（降价 90%）
- **GLM-Z1-Air**：¥50/亿 tokens
- **GLM Coding Plan**：订阅套餐 ¥20/月起

适合场景

- 预算极度有限的项目（FlashX 超低价）
- 中文学术研究
- 代码生成（Coding Plan）
- 需要超长上下文的应用（GLM-4-Long）

中国平台特殊优势

- 清华背景，技术实力强
- 定价全行业最低（FlashX）
- 免费模型性能优秀
- 支持超长文本处理

18. 月之暗面（Kimi）

核心模型

- **Kimi K2 Thinking**（2025年11月6日发布，最新！）
  - Kimi系列能力最强的开源思考模型
  - 第一代原生支持边思考边使用工具的Thinking Agent
  - 深度思考与工具编排完美融合
  - Humanity's Last Exam（HLE）44.9%，超越GPT-5、Grok-4、Claude 4.5
  - 训练成本仅460万美元（CNBC报道）
- **Kimi K2**（2025年7月11日，突破性开源模型）
  - 1万亿总参数，320亿激活参数
  - Agent能力专长
- **moonshot-v1-8k / 32k / 128k**（基础系列）
- **moonshot-v1-vision-preview**（多模态）

注册流程

1. 访问 https://platform.moonshot.cn
2. 注册账号（手机号/邮箱）
3. 在 API Keys 页面 创建密钥

免费额度与价格

- **免费额度**：新用户赠送 ¥15 额度
- **moonshot-v1-8k**：¥12/百万 tokens
- **moonshot-v1-128k**：¥60/百万 tokens
- **Kimi K2**：
  - 输入（Cache Miss）：$0.60/百万 tokens
  - 输入（Cache Hit）：$0.15/百万 tokens
  - 输出：$2.50/百万 tokens
- **Vision 系列**：同文本模型价格

适合场景

- 超长上下文需求（最高 200K 中文字符）
- 文档分析、代码审查
- 研究论文阅读与总结
- Agent 开发（K2 模型专长）

中国平台特殊优势

- 行业领先的超长上下文能力
- 强大的 Agent 能力
- 兼容 OpenAI 和 Anthropic API 格式
- 适合处理大型文档

19. MiniMax

核心模型

- **MiniMax M2**（2025年10月发布，开源）
- **abab6.5s**（主力模型）
- **abab6.5t**（文本理解）

注册流程

1. 访问 https://www.minimaxi.com
2. 注册并完成实名认证
3. 在控制台创建 API Key

免费额度与价格

- **免费额度**：完成认证后赠送 1 亿 tokens（可用于 abab6.5s）
- **abab6.5s**：¥5/百万 tokens
- **MiniMax M2**：
  - 输入：$0.3/百万 tokens（¥2.1/百万 tokens）
  - 输出：$1.2/百万 tokens（¥8.4/百万 tokens）
  - 仅为 Claude Sonnet 价格的 8%
- **M2 Agent 平台**：暂时免费

适合场景

- 音视频生成（MiniMax 专长）
- 多模态应用
- 预算有限的创业项目
- 需要开源模型的企业

中国平台特殊优势

- M2 模型开源且性价比极高
- 强大的语音、视频生成能力
- 免费额度慷慨

- 适合多媒体应用开发

20. 字节豆包

核心模型

- **Doubao-pro-32k**
- **Doubao-lite-32k**
- **Doubao Image**（图像生成）

注册流程

1. 访问 https://www.volcengine.com/products/doubao
2. 注册火山引擎账号（需实名认证）
3. 开通豆包服务
4. 创建 API Key

免费额度与价格

- **免费额度**：50 万 tokens/月，可领取额外 5 亿 tokens
- **Doubao-lite-32k**：
  - 输入：¥0.3/百万 tokens
  - 输出：¥0.6/百万 tokens
- **Doubao-pro-32k**：
  - 输入：¥0.8/百万 tokens
  - 输出：¥2/百万 tokens

适合场景

- 字节跳动生态集成（抖音、今日头条）
- 内容推荐、审核
- 中等规模应用
- 需要稳定性的商业项目

中国平台特殊优势

- 字节跳动技术支持
- 与抖音、今日头条深度集成
- 价格适中，性价比高
- 稳定可靠的服务

21. 讯飞星火

核心模型

- **Spark X1 升级版**（2025年7月，最新！）
  - 深度推理模型

    - 多语言能力扩展至130+语言
- **Spark 4.0 Ultra**
    - 支持内置插件：搜索、天气、日期、诗词、成语、股票
    - 支持在线搜索检索返回源标题和地址
    - 支持System和Function Call功能
- **Spark Max / Pro / Lite**
    - 六个版本：Lite、Pro、Pro-128K、Max、Max-32K、4.0 Ultra
    - Spark Lite：永久免费

注册流程

1. 访问 https://xinghuo.xfyun.cn
2. 注册讯飞账号（需实名认证）
3. 开通星火服务
4. 在 开放平台 创建应用并获取 API Key

免费额度与价格

- **Spark Lite**：永久免费
- **个人用户**：200 万 tokens/年（免费额度）
- **Spark 3.5 Max**：¥0.21/万 tokens（仅为百度、阿里的 1/5）
- **Spark X1**：已上线开放平台

适合场景

- 语音识别与合成（讯飞专长）
- 教育、医疗等垂直领域
- 需要永久免费模型的项目
- 多模态应用（语音+文本）

中国平台特殊优势

- 行业领先的语音技术
- Lite 版永久免费
- 价格极具竞争力（Max 版）
- 适合教育、医疗等场景

国际平台横向对比表（2025 Q4）

平台	免费额度	最佳模型	输入价格	输出价格	延迟	注册难度	适合人群
OpenAI	GPT-4o mini 免费层	GPT-5	$5/1M	$15/1M	⭐⭐⭐⭐⭐	⭐⭐⭐	商业产品
Anthropic	$5	Claude Sonnet 4.5	$3/1M	$15/1M	⭐⭐⭐⭐	⭐⭐⭐	企业/研究
Google AI	15 RPM 免费	Gemini 2.5 Pro	$3.5/1M	$10.5/1M	⭐⭐⭐⭐	⭐⭐	Google生态

平台	免费额度	最佳模型	输入价格	输出价格	延迟	注册难度	适合人群
Together AI	$25	Llama 4 Maverick	$0.10/1M	-	⭐⭐	⭐	学生/研究者
Fireworks AI	$1	Llama 4 Maverick	$0.9/1M	-	⭐⭐⭐⭐⭐	⭐	生产部署
Perplexity	Pro 用户 $5/月	Sonar Large	$5/1M	-	⭐⭐⭐	⭐⭐	联网搜索
Mistral AI	$500	Mistral Large 2	$2/1M	$6/1M	⭐⭐⭐	⭐⭐	欧洲/预算
OpenRouter	$1	聚合所有	+5.5%费用	+5.5%费用	取决底层	⭐	快速实验
Groq	免费层	Llama 4 Maverick	竞争力强	竞争力强	⭐⭐⭐⭐⭐	⭐	低延迟
Replicate	无	各类生成模型	按秒计费	按秒计费	⭐⭐⭐	⭐	图像/视频
Hugging Face	每月免费额度	100万+开源模型	PRO $9/月	-	⭐⭐	⭐	开源实验
Cohere	5000单位/月	Command A 03-2025	$0.15/1M	$0.60/1M	⭐⭐⭐	⭐	企业NLP
AI21 Labs	$10（3个月）	Jamba 1.7	$0.0188/1M	-	⭐⭐⭐	⭐⭐	任务型API

中国平台横向对比表（2025 Q4）

平台	免费额度	最佳模型	价格（¥/百万tokens）	实名认证	支付方式	适合场景
阿里云百炼	100万tokens	Qwen3-Max Thinking	¥40（Max）	需要	支付宝/微信	电商/客服
腾讯混元	100万tokens	Hunyuan Large	**开源**	需要	微信/支付宝	社交/游戏
百度千帆	ERNIE 3.5免费	ERNIE 5.0-Preview	**即将发布**	需要	支付宝/微信	搜索/内容
智谱AI	Flash免费	GLM-4.6	¥0.8（输入）	可选	支付宝/微信	学术/代码
月之暗面	¥15	Kimi K2 Thinking	¥4.2（输入）	可选	支付宝/微信	长文档/Agent
MiniMax	1亿tokens	MiniMax M2	¥2.1（输入）	需要	支付宝/微信	音视频

平台	免费额度	最佳模型	价格（¥/百万tokens）	实名认证	支付方式	适合场景
字节豆包	50万tokens	Doubao-pro-32k	¥0.8（输入）	需要	支付宝/微信	内容推荐
讯飞星火	200万tokens/年	Spark 3.5 Max	¥0.21/万tokens	需要	支付宝/微信	语音/教育

价格说明：

- 国际平台价格以美元/百万tokens计
- 中国平台价格以人民币/百万tokens计
- 延迟评级：⭐（较慢）到 ⭐⭐⭐⭐⭐（极快）
- 注册难度：⭐（简单）到 ⭐⭐⭐（复杂）

选型建议：根据你的需求选择平台

按身份选择

在校大学生 / 课程项目

国际平台推荐：

- **首选**：Together AI（25免费额度）或$MistralAI$（500 免费额度）
- **备选**：Google Gemini（15 RPM 免费层，可商用）、Hugging Face（开源实验）

中国平台推荐：

- **首选**：智谱 AI（GLM-4-Flash 免费）或讯飞星火（Spark Lite 永久免费）
- **备选**：百度千帆（ERNIE 3.5 免费）、MiniMax（1亿tokens免费）

AI 科技博主 / 开源贡献者

国际平台推荐：

- **首选**：OpenRouter（便于多模型对比，支持 Referer 标注）
- **备选**：Hugging Face（开源社区）、Together AI（开源模型全）

中国平台推荐：

- **首选**：智谱 AI（价格最低，适合大量测试）
- **备选**：月之暗面（技术前沿）、MiniMax（M2 开源）

商业产品 MVP

国际平台推荐：

- **首选**：GPT-5（OpenAI）或 Claude Sonnet 4.5（Anthropic）
- **理由**：2025年最强模型，稳定性与文档最佳，GPT-5降低45%事实错误

中国平台推荐：

- **首选**：ERNIE 5.0（百度，即将发布）或 Qwen3-Max（阿里）
- **理由**：国内顶尖模型，大厂背书，服务稳定，符合国内合规

研究实时信息应用

国际平台推荐：

- **唯一选择**：Perplexity（Sonar 模型联网能力不可替代）

部署高性能生产服务

国际平台推荐：

- **首选**：Fireworks AI（最低延迟与 SLA 保障）或 Groq（超高速推理）

中国平台推荐：

- **首选**：字节豆包或腾讯混元
- **理由**：大厂技术支持，稳定性高

按场景选择

预算极度有限（免费/超低价）

国际平台：

- Google Gemini（15 RPM 免费层）
- Mistral AI（$500 免费额度）
- Together AI（$25 免费额度）
- Hugging Face（PRO $9/月）

中国平台：

- 智谱 AI GLM-4-FlashX（¥10/亿tokens，全行业最低）
- 讯飞星火 Lite（永久免费）
- 百度千帆 ERNIE 3.5（免费）
- MiniMax M2（¥2.1/百万输入tokens）

需要超长上下文（100K+ tokens）

国际平台：

- Google Gemini 1.5 Pro（1M tokens）
- Claude 3.5 Sonnet（200K tokens）
- Together AI Llama 3.1（128K tokens）

中国平台：

- 月之暗面 Kimi（200K 中文字符）
- 智谱 GLM-4-Long（超长上下文）

- 腾讯混元 Standard-256K

多模态应用（图像+文本）

国际平台：

- OpenAI GPT-4o（最强多模态）
- Google Gemini 2.5 Flash Image
- Mistral Pixtral 12B（免费）

中国平台：

- 阿里云 Qwen-VL
- 月之暗面 Vision 系列
- MiniMax M2

代码生成与编程

国际平台：

- Claude Sonnet 4.5（世界最佳编程模型，2025年9月）
- GPT-5（SWE-bench 74.9%）
- Mistral Codestral 25.01
- Together AI Llama 4 Behemoth（超越GPT-4.5）

中国平台：

- 智谱 GLM-4.6（代码能力对齐Claude Sonnet 4，国产最强）
- 阿里云 Qwen3（世界最强开源）
- 月之暗面 K2 Thinking（Agent + 代码专长）

语音/音频处理

中国平台：

- 讯飞星火（行业领先的语音技术）
- MiniMax（音视频生成）

图像/视频生成

国际平台：

- Replicate（Veo 2, SDXL, Kling 等）
- Hugging Face（Stable Diffusion 等）

中国平台：

- MiniMax（视频生成专长）
- 字节豆包（Doubao Image）

针对中国开发者的专门建议

网络访问考虑

- **国际平台**：需要稳定的网络环境访问，可能需要使用代理
- **中国平台**：直接访问，速度快，无需额外配置

支付方式

- **国际平台**：通常需要国际信用卡（Visa/Mastercard），部分支持 PayPal
- **中国平台**：支持支付宝、微信支付等本地支付方式，更便捷

语言能力

- **国际平台**：英文能力较强，中文理解相对较弱（除 Qwen 等）
- **中国平台**：中文理解和生成能力普遍优于国际平台，更适合中文应用

合规性

- **国际平台**：数据可能存储在海外，需考虑数据出境合规问题
- **中国平台**：符合国内数据安全法规，适合政府、金融、医疗等敏感领域

推荐组合策略

个人开发者：

- 主力：智谱 AI（超低价）+ 讯飞星火 Lite（免费）
- 备用：Google Gemini（免费层）或 OpenRouter（模型对比）

创业团队：

- 中文业务：阿里云 Qwen3-Max 或百度 ERNIE 5.0（即将）
- 国际业务：GPT-5 或 Claude Sonnet 4.5
- 成本优化：智谱 AI GLM-4.6 + MiniMax M2

企业用户：

- 主力：阿里云/腾讯云/百度云（根据现有生态选择）
- 备用：OpenAI 或 Anthropic（高端需求）
- 特殊场景：讯飞星火（语音）、Perplexity（联网搜索）

结语

AI API 平台已从"单极垄断"走向"百花齐放"，从国际巨头到中国本土，从通用模型到垂直场景，开发者有了前所未有的选择空间。特别是2025年下半年，AI领域迎来了模型能力的集中爆发期，多个重磅模型相继发布。理解每个平台的定位、模型优势、注册门槛与定价策略，是你迈向高效 AI 开发的关键一步。

🚀 2025年11月最新突破（重点关注）

国际平台重大更新：

- **GPT-5**（OpenAI，2025年8月）：全球最强统一模型，降低45-80%错误率，数学94.6%、编程74.9%
- **Claude Sonnet 4.5**（Anthropic，9月29日）：世界最佳编程模型，最强Agent构建能力
- **Gemini 3.0**（Google，即将11月）：预告更强AI Agent能力，1M上下文
- **Llama 4系列**（Meta，4月）：10M超长上下文（Scout），MoE架构创新

中国平台惊艳表现：

- **ERNIE 5.0-Preview**（百度，11月8日）：LMArena全球第二、国内第一，超越GPT-5-High
- **Kimi K2 Thinking**（月之暗面，11月6日）：最强Thinking Agent，训练成本仅460万美元，HLE 44.9%
- **Hunyuan Large**（腾讯，11月5日）：MoE架构完全开源，企业可自由部署
- **Qwen3-Max Thinking**（阿里，11月2日）：1T参数深度思考模式，AIME 25达100%
- **GLM-4.6**（智谱，9月30日）：国产代码能力最强，对齐Claude Sonnet 4

🏆 性价比之王推荐

- **国际免费最高**：Mistral AI（500）、$TogetherAI$（25）
- **国内超低价**：智谱 GLM-4-FlashX（¥10/亿tokens）
- **永久免费**：讯飞星火 Lite、腾讯混元 Lite、百度 ERNIE 3.5/4.0（网页版）
- **开源之王**：腾讯混元 Large（MoE完全开源）、Qwen3系列（Apache 2.0）、Kimi K2（1T参数开源）

📋 核心建议

1. **追踪最新模型**：2025年模型更新频繁，建议关注本文档持续更新（v1.1已更新至11月）
2. **从免费开始**：新手优先选择免费额度慷慨的平台（Mistral AI 500、$TogetherAI$25、智谱 GLM-4-Flash 免费）
3. **多平台测试**：使用 OpenRouter 或自行对比，找到最适合你业务的模型
4. **关注性价比**：不要盲目追求最强模型，智谱 GLM-4-FlashX（¥10/亿tokens）等超低价模型可能更适合你
5. **考虑生态**：如果已在使用阿里云/腾讯云/Google Cloud，优先考虑同生态的 AI 服务
6. **合规优先**：中国企业应优先考虑本土平台，确保数据合规
7. **Agent优先**：如需构建复杂Agent，关注Claude Sonnet 4.5、Kimi K2 Thinking、Qwen3-Max Thinking

> 🌐 建议行动：
> 选一个平台，今天就注册并跑通你的第一个 `Hello, LLM!` 请求。
> 代码的世界，从一次 API 调用开始。

附录：各平台官网速查

国际平台

- OpenAI：https://platform.openai.com
- Anthropic：https://console.anthropic.com
- Google AI：https://aistudio.google.com
- Together AI：https://together.ai
- Fireworks AI：https://fireworks.ai
- Perplexity：https://www.perplexity.ai
- Mistral AI：https://console.mistral.ai
- OpenRouter：https://openrouter.ai
- Groq：https://console.groq.com

- Replicate：https://replicate.com
- Hugging Face：https://huggingface.co
- Cohere：https://cohere.com
- AI21 Labs：https://www.ai21.com

中国平台

- 阿里云百炼：https://bailian.console.aliyun.com
- 腾讯混元：https://cloud.tencent.com/product/hunyuan
- 百度千帆：https://cloud.baidu.com/product/wenxinworkshop
- 智谱 AI：https://open.bigmodel.cn
- 月之暗面：https://platform.moonshot.cn
- MiniMax：https://www.minimaxi.com
- 字节豆包：https://www.volcengine.com/products/doubao
- 讯飞星火：https://xinghuo.xfyun.cn

版本历史：

- v1.1（2025-11-12晚）：**重大更新** - 更新所有平台最新模型信息至2025年11月
  - 国际平台：GPT-5、Claude 4系列、Gemini 2.5/3.0、Llama 4系列、Magistral等
  - 中国平台：Qwen3-Max Thinking、ERNIE 5.0-Preview、GLM-4.6、Hunyuan Large、Kimi K2 Thinking等
- v1.0（2025-11-12）：全面更新，新增 13 个平台（5 个国际 + 8 个中国），更新所有定价信息至 2025 年 11 月

反馈与贡献：
如需获取本文的 Markdown 源码、配套代码模板，或希望补充更多平台信息，欢迎提出建议！